CLAIMS

What is claimed is:

1. A normalization system, comprising:

an interface component that receives data corresponding to a heterogeneous knowledge base; and

a normalization component that applies a model that predicts accuracy or quality of results to provide a regularized understanding of the knowledge base.

2. The system of claim 1, the interface component processes questions posed by users.

3. The system of claim 1, the utility model dynamically controls extraction of previously unknown or disassociated information from the knowledge base.

4. The system of claim 1, the utility model controls a number of queries submitted to the knowledge base given decision-theoretic considerations.

5. The system of claim 1, the knowledge base includes at least one a local database, a file, a directory, an electronic encyclopedia, a dictionary, a remote database, and a remote web site.

6. The system of claim 1, where the knowledge base is the World Wide Web.

7. The system of claim 1, the utility model applies a cost-benefit analysis to dynamically control the number and types of attempts made to acquire information or answers from the knowledge base in response to a question or questions.

8. The system of claim 7, the utility model includes an analysis of the costs of searching for information versus the benefits or value of obtaining more accurate answers to questions.

9. The system of claim 1, the interface component initiates a dialog with users based upon predetermined probability thresholds or other criteria that includes a cost-benefit analyses that considers when it would be best to ask a user to reformulate a question rather than expending effort on processing a query that may be expensive in terms of searching for information from the knowledge base or likely to yield inaccurate results.

10. The system of claim 9, the dialog is initiated from an assessment of a cost of delay and effort associated with a query reformulation and a likelihood that a reformulation would lead to an improved result.

11. The system of claim 1, further comprising a preference component that enables users to assess or select various parameters that influence the utility model.

12. The system of claim 11, the preference component processes at least one of a user setting for a cost, a value, and a language preference.

13. The system of claim 12, the preference component includes a model where a user assesses a parameter *v*, indicating a dollar value of receiving a correct answer to a question, and where a parameter *c* represents a cost of each query rewrite submitted to a search engine.

14. The system of claim 13, further comprising a value of receiving an answer expressed as a function of details of a current context, the value of the answer is linked to at least one of a type of question, an informational goal, and a time of day for a user.

15. The system of claim 13, further comprising determining a cost of submitting queries as a function of at least one of a current load sensed on a search engine or the numbers of queries being submitted by a user's entire organization to a third-party search service.

16. The system of claim 15, further comprising determining the costs non-linearly with increasing numbers of queries.

17. The system of claim 16, further comprising assessing costs of n at a first level of cost by a search service supporting a question-answering system at an enterprise, after which expenses are incurred in a supra-linear manner, n being an integer.

18. The system of claim 1, the utility model determines an expected value of submitting n queries as a product of a likelihood of an answer, given evidence E about a query, $p(A|E,n,\xi)$, and a value of obtaining a correct answer v, $p(A|E,n,\xi)\ v$.

19. The system of claim 18, the value of an answer, v, is assessed in terms of the cost of queries, c as a multiple k of the cost of each query c, $v=kc$.

20. The system of claim 19, further comprising a cost model that grows linearly with a number of queries, nc.

21. The system of claim 20, further comprising a utility model that determines an ideal number of queries to submit, optimizes a net expected value, computed as a difference of the expected value and cost, for different n.

22. The system of claim 21, the ideal number of queries is expressed as:

$$n^{*}= \arg\max_{n} p(A|E,n,\xi)kc - nc.$$

23. A computer readable medium having computer readable instructions stored thereon for implementing the interface component and the normalization component of claim 1.

24. A method to normalize a database, comprising:
automatically forming a set of queries from a question posed by a user; and
performing a cost-benefit analysis on the set of queries to generate a query subset.

25. The method of claim 24, further comprising automatically ranking the set of queries in an order of likelihood of providing a suitable answer.

26. The method of claim 24, further comprising automatically training at least one model to for the query subset.

27. The method of claim 24, further comprising submitting the query subset to at least one search engine.

28. The method of claim 27, further comprising receiving results from the at least one search engine and automatically composing an answer.

29. A system to facilitate database normalization, comprising:
means for formulating a query set from a user question; and
means for forming a query subset from the query set based at least in part on a utility model employed for normalizing the database.

30. A question-answering system, comprising:

a rewriting component that receives a user query and automatically formulates a set of queries; and

a cost-benefit component to reduce the set of queries based upon an analysis of expected gains in accuracy of an answer in view of associated costs for additional queries.

31. The system of claim 30, further comprising a ranking component to determine an ordering for the set of queries.

32. The system of claim 30, further comprising an answer composer to formulate an answer from the reduced set of queries.

33. The system of claim 30, further comprising at least one search engine to process the set of queries.

34. The system of claim 30, further comprising a component to process full text and/or text summaries of articles returned by a search engine.

35. The system of claim 30, further comprising a component that learns logical or statistical predictive models that predict an accuracy or quality of answers as a function of nature or number of queries issued to a knowledge base.

36. The system of claim 30, further comprising a component that learns logical or statistical predictive models that predict an age appropriateness of answers as a function of nature or number of queries issued to a knowledge base.

37. The system of claim 35, the knowledge base includes at least one search engine.

38. The system of claim 35, the models employ Bayesian learning procedures.

39. The system of claim 30, further comprising a component to analyze at least one feature, the feature including at least one of a conjunctional and a phrasal rewrite.

40. The system of claim 39, the feature includes at least one of a longest phrase, a word length, a number of capital letters, a number of phrases, a number of stop words, a number of words, a percentage of stop words, a number of primary parses, a number of secondary parses, and a measure of grammatical suitability.

41. The system of claim 39, further comprising higher-level features including a distribution of topics associated with results of queries.

42. The system of claim 41, the features are identified with a statistical classifier that assigns topics based on text being analyzed.,

43. The system of claim 39, further comprising tags that are derived from natural-language parses of initial questions, and text of results or text of snippets returned from the results.

44. The system of claim 39, further comprising a component to derive higher-level informational goals of a user, as derived from assessing goals directly, or as inferred from an analysis of the user's initial question.

45. The system of claim 39, further comprising classes of features including attributes and statistics of attributes of morphological or semantic aspects of an initial (1) question and/or (2) one or more query results.

46. The system of claim 45, the classes include at least one of words and phrases, parts of speech, structure of natural language parse, length, topics and distribution of topics, and inferred or assessed informational goals or intentions.

47. A normalization system, comprising:

an interface component that receives data corresponding to a heterogeneous knowledge base; and

a normalization component that applies a model that predicts accuracy or quality of results in conjunction with a utility model to provide a regularized understanding of the value of performing different information extraction actions from the knowledge base.

48. The system of claim 47, the interface component processes questions posed by users.

49. The system of claim 47, further comprising a dialog component that makes a decision when to engage a user to request a reformulated question or additional information.

50. The system of claim 49, the dialog component alerts the user about the cost of receiving a good answer, or recommends that a query be attempted elsewhere.